**From:** rkjohnso@unlnotes.unl.edu%inter2 [mailto:rkjohnso@unlnotes.unl.edu]
**Sent:** Wednesday, November 30, 2005 12:39 PM
**To:** OConnor, Martin
**Subject:**

Martin E. O'Connor
Chief
Standardization Branch
Livestock and Seed Program
Agricultural Marketing Service
USDA
1400 Independence Avenue, S.W.
Room 2607-S
Washington, DC 20250

I will not be able to attend the hearing on Dec 6 during which testimony regarding DNA testing for breed certification will be taken, but I have some thoughts on the topic that I would like you to consider.

I believe that DNA tests could some day be developed to certify breed purity, but with current information there is considerable room for error. The reason is that some breeds are fixed for certain markers, whereas other breeds are segregating for these same markers, so it will be very difficult to achieve 100% accuracy of DNA tests. Therefore, a certification program must first define a level of accuracy acceptable to all parties - those whose product is being certified as well as the purchaser of the product. Two types of errors must be considered - these include situations in which some purebreds do not pass the test and situations in which non-purebreds pass the test.

A scientific process is needed to establish panels of markers for certification. Breeds must be characterized for frequencies of alleles of a farily large panel of DNA markers. There are some logical markers, such as genes coding for color, to include in the panel, but many more are needed as a small panel of markers will be insufficient to characterize all breeds. Then a panel of markers can be developed that distinguish among pairs of breeds with known levels of accuracy. Different panels will likely be needed to distinguish different pairs of breeds. Then the question of how to differentiate crossbred and composite populations from the purebreds must be addressed. I don't think crossbreds and purebreds can be accurately differentiated with current information.

Careful thought must be given to long-term consequences of a certification standard. Once panels of markers are accepted as standards for certification, it would be relatively easy for a creative geneticist to establish composite populations that could pass the test.

In summary, it seems that current information on marker frequencies across breeds of livestock are insufficient to establish a reliable certification program and to asses whether crossbreds can be distinguished from purebreds. My thoughts on the topic are laid out more fully in the attached paper.

Thanks for considering my input.

Dr. Rodger Johnson
Omtvedt Professor
Animal Science Department
218 Animal Science
University of Nebraska
Lincoln, NE 68583-0908
Tel.: 402/472-6404
Fax: 402/472-6362
email: Rjohnson5@unl.edu

# Some thoughts on the difficulty of certifying breeds with molecular markers

Rodger Johnson
Animal Science Department
University of Nebraska

How many markers are necessary to certify breed purity from DNA tests? The answer is it depends. It depends on the number of breeds. It depends on whether the goal is to differentiate only among purebreds or also between purebreds and crossbreds. But most of all it depends on the frequency of the markers within each breed. Below, I examine these issues with examples. An assumption in these examples is that these markers are not being selected for and are not linked with genes that are being selected. This may be an oversimplification, but space does not allow the selection model to be addressed.

**The single locus example**: Assume several possible alleles of a single DNA marker, denoted as the A locus, and all breeds are fixed for a different allele. Genotypes for Breeds X, Y, and Z would then be:

| Breed | Genotype |
|---|---|
| X | $A_1A_1$ |
| Y | $A_2A_2$ |
| Z | $A_3A_3$ |

A simple one-marker DNA test, assuming it is 100% accurate, will differentiate among breeds without error. Furthermore, all first cross progeny ($F_1$) can be differentiated without error. X by Y $F_1$ are all $A_1A_2$, X by Z are $A_1A_3$, and Y by Z are $A_2A_3$.

So far, so good. The simple test differentiates among all breeds and crosses with 100% accuracy. But now suppose the hybrids are mated together, for example an XY hybrid with another XY hybrid, or an XY hybrid with an XZ hybrid. Expected genotypic distributions among progeny are:

| Parents | $A_1A_1$ | $A_1A_2$ | $A_1A_3$ | $A_2A_2$ | $A_2A_3$ | $A_3A_3$ |
|---|---|---|---|---|---|---|
| XY x XY | .25 | .50 | | .25 | | |
| XY x XZ | .25 | .25 | .25 | | .25 | |

The DNA test now is unreliable. In the first case, the DNA test would certify 25% of the progeny as Breed X and 25% as Breed Y. In the second case, 25% would certify as Breed X and the only certainty is that none of the progeny are Breeds Y or Z.

The above example is much oversimplified. There are likely to be very few markers, if any, for which all breeds are fixed for different alleles. Markers likely exist for which one breed is fixed (completely homozygous) for a particular allele, but other breeds are likely to be segregating at this marker. For example, assume Breed X is fixed for allele $A_1$, Breed Y has allele frequencies of .3 $A_1$ and .7 $A_2$, and Breed Z, although it has a high

frequency of allele $A_3$, is segregating for all three alleles with frequencies .05 $A_1$, .1 $A_2$, and .85 $A_3$. The expected distribution of progeny in the breeds is:

| Breed | $A_1A_1$ | $A_1A_2$ | $A_1A_3$ | $A_2A_2$ | $A_2A_3$ | $A_3A_3$ |
|---|---|---|---|---|---|---|
| X | 1 | | | | | |
| Y | .09 | .42 | | .49 | | |
| Z | .0025 | .01 | .085 | .01 | .17 | .7225 |

There is now no 100% reliable test for any breed. If animals with genotype $A_1A_1$ only are certified as Breed X, then Breed X animals are never misclassified, but 9% of Breed Y progeny and .25% of Breed Z progeny also certify as Breed X. Certification of animals as Breeds X or Z is even less reliable.

**Multiple markers**. With multiple markers, determining the number of markers needed to correctly classify breeds is an even larger statistical problem. The number of possible genotypes increases greatly with increases in the number of markers and alleles per marker, making illustration of more complex genotypic models difficult. The complexity of the problem is illustrated in Table 1 that shows expected distributions of combinations of genotypes for a three-marker model with two alleles at each marker.

Assume that each breed has a high frequency of one allele at a particular marker, with the other breeds segregating at intermediate frequencies for that marker. Breed X has a high frequency (.9) of allele A2 at marker A, whereas Breeds Y and Z have intermediate frequencies (.3 and .6, respectively). Breed Y is distinguished by an especially high frequency of allele B2 at marker B, and Breed Z is fixed for allele C1 at marker C.

There is no clear picture about choosing genotypes to classify breeds. Although some genotypes are rare in certain breeds, no genotype across markers differentiates breeds with 100% accuracy. The only decision made with 100% accuracy is that Breed Z animals cannot have the C2 allele, but 64% of Breed X and 25% of Breed Y also pass that test. On the surface, it seems that using the most common genotype of each breed as the classification criterion may differentiate breeds. Below are the genotypes that are most common in each breed, and the proportion of animals of the other breeds that also have those genotypes. The most common genotype in Breed Y is very rare in the other breeds, but 78% of Breed Y animals also would fail the test by that criterion, certainly an unacceptable outcome. Extending this argument to most common genotypes does not add clarity to the picture.

| | | Frequency in each breed | | |
|---|---|---|---|---|
| | Most common genotype | X | Y | Z |
| Breed X | A2A2B2B2C1C1 | .254 | .02 | .176 |
| Breed Y | A1A1B2B2C1C2 | .0015 | .22 | 0 |
| Breed Z | A1A2B2B2C1C1 | .056 | .095 | .235 |

| Table 1. Expected distribution of genotypes across three marker loci in each of three breeds | | | | | |
|---|---|---|---|---|---|
| | | Marker frequencies | | | |
| | | | Breed | | |
| | | Marker | X | Y | Z |
| | | A1 | 0.1 | 0.7 | 0.4 |
| | | A2 | 0.9 | 0.3 | 0.6 |
| | | B1 | 0.3 | 0.05 | 0.3 |
| | | B2 | 0.7 | 0.95 | 0.7 |
| | | C1 | 0.8 | 0.5 | 1 |
| | | C2 | 0.2 | 0.5 | 0 |
| Genotype | | | Genotypic frequencies | | |
| Marker A | Marker B | Marker C | X | Y | Z |
| A1A1 | B1B1 | C1C1 | 0.000576 | 0.000306 | 0.0144 |
| A1A1 | B1B1 | C1C2 | 0.000288 | 0.000613 | 0 |
| A1A1 | B1B1 | C2C2 | 0.000036 | 0.000306 | 0 |
| A1A1 | B1B2 | C1C1 | 0.002688 | 0.011638 | 0.0672 |
| A1A1 | B1B2 | C1C2 | 0.001344 | 0.023275 | 0 |
| A1A1 | B1B2 | C2C2 | 0.000168 | 0.011638 | 0 |
| A1A1 | B2B2 | C1C1 | 0.003136 | 0.110556 | 0.0784 |
| A1A1 | B2B2 | C1C2 | 0.001568 | 0.221113 | 0 |
| A1A1 | B2B2 | C2C2 | 0.000196 | 0.110556 | 0 |
| A1A2 | B1B1 | C1C1 | 0.010368 | 0.000263 | 0.0432 |
| A1A2 | B1B1 | C1C2 | 0.005184 | 0.000525 | 0 |
| A1A2 | B1B1 | C2C2 | 0.000648 | 0.000263 | 0 |
| A1A2 | B1B2 | C1C1 | 0.048384 | 0.009975 | 0.2016 |
| A1A2 | B1B2 | C1C2 | 0.024192 | 0.01995 | 0 |
| A1A2 | B1B2 | C2C2 | 0.003024 | 0.009975 | 0 |
| A1A2 | B2B2 | C1C1 | 0.056448 | 0.094763 | 0.2352 |
| A1A2 | B2B2 | C1C2 | 0.028224 | 0.189525 | 0 |
| A1A2 | B2B2 | C2C2 | 0.003528 | 0.094763 | 0 |
| A2A2 | B1B1 | C1C1 | 0.046656 | 5.63E-05 | 0.0324 |
| A2A2 | B1B1 | C1C2 | 0.023328 | 0.000113 | 0 |
| A2A2 | B1B1 | C2C2 | 0.002916 | 5.63E-05 | 0 |
| A2A2 | B1B2 | C1C1 | 0.217728 | 0.002138 | 0.1512 |
| A2A2 | B1B2 | C1C2 | 0.108864 | 0.004275 | 0 |
| A2A2 | B1B2 | C2C2 | 0.013608 | 0.002138 | 0 |
| A2A2 | B2B2 | C1C1 | 0.254016 | 0.020306 | 0.1764 |
| A2A2 | B2B2 | C1C2 | 0.127008 | 0.040613 | 0 |
| A2A2 | B2B2 | C2C2 | 0.015876 | 0.020306 | 0 |

Summary: The simple examples illustrate that defining the number of DNA markers necessary to accurately classify animals according to breed depends greatly on knowledge of allele frequencies at each marker within all breeds of interest. Marker data for some genes to begin determining this number may be available for certain breeds, but

it is unlikely that adequate data exist for all breeds. Until proper sampling within breeds occurs, considerable room for error exists. Furthermore, identifying markers to distinguish crossbred animals or animals from composite populations from those in foundation pure breeds seems to be nearly an impossible task. Depending on the level of accuracy desired, two kinds of misclassifications are possible. If the criteria are too stringent, true purebreds may be excluded because they have a rare genotype within the breed. If the policy is more liberal, some animals of other breeds may be incorrectly classified and certainly animals from crossbred or composite populations can be erroneously classified.

A scientific approach to choosing markers for breed certification is recommended. Possible steps include:

1. Define the desired level of precision with which allele frequencies are to be estimated. This value is known as the standard error of estimated allele frequency.
    a. The standard error (SE) of an estimated allele frequency is $\sqrt{pq/2N}$, where p and q are estimates of allele frequencies in the sample and N is sample size, the number of randomly chosen animals of the population.
    b. For p = .95 and q = .05, a sample size of 100 produces a standard error of estimate of .0154. The probability is approximately .95 that the population value of p, the frequency of the major allele, is in the interval of p ± 2SE; in this example, the interval is .919 to .981.
    c. SE of estimates increase as allele frequencies approach .5, in which case a sample of 100 produces an SE of .0354.
2. Sample all populations so as to achieve the desired level of precision. This can be accomplished in two ways
    a. Draw a random sample of the targeted number from the entire population.
    b. Because current sires and dams are expected to have similar allele frequencies, the next generation of progeny will have allele frequencies similar to those of their sires and their dams. An acceptable sampling process seems to be to sample all current sires. Allele frequencies could then be calculated from an average value weighted by the number of litters sires are expected to produce.
3. Carefully genotype the sample of animals from all populations to characterize their allele frequencies and expected genotypic frequencies.
    a. Carefully analyze these data to select a panel of markers that accurately certifies breed of origin. It is likely that the markers and the number required will be different for different breed contrasts. For example, one panel of n markers may differentiate Breed X from Breed Y, but a different panel of m markers may be required to differentiate Breed X from Breed Z.
4. Some miscalculation seems inevitable, either due to genotyping errors or to too liberal or too conservative policies. Use the data to define acceptable error rates for both types of miscalculation.